# *CRITERIA FOR CHOOSING THE BEST NEURAL NETWORK:*

# *PART I*

*J. E. Angus*

DTIC
ELECTE
MAR 23 1992
D

*Report No. 91-16*

92-07423

Approved for public release: distribution unlimited.

NAVAL HEALTH RESEARCH CENTER
P.O. BOX 85122
SAN DIEGO, CALIFORNIA 92186-5122

NAVAL MEDICAL RESEARCH AND DEVELOPMENT COMMAND
BETHESDA, MARYLAND

# CRITERIA FOR CHOOSING THE BEST NEURAL NETWORK: PART I

John E. Angus

Department of Mathematics
The Claremont Graduate School
143 E. Tenth Street
Claremont, CA 91711-3988

and

Medical Information Systems and Operations Research Department
Naval Health Research Center
P.O. Box 85122
San Diego, CA 92186-5122

September 1991

Report No. 91-16, supported by the Naval Health Research Center under the American Society for Engineering Education Navy Summer Faculty Research Program. The views expressed in this report are those of the author and do not reflect the official policy or position of the Department of Defense, nor the U.S. Government. Approved for public release, distribution unlimited.

# CRITERIA FOR CHOOSING THE BEST NEURAL NETWORK
# PART I

John E. Angus
Department of Mathematics
The Claremont Graduate School
Claremont, CA 91711

## ABSTRACT

This paper considers the problem of determining a parsimonious neural network for use in prediction/generalization based on a given fixed learning sample. Both the classification and nonlinear regression contexts are addressed. Following an introduction to the problem and survey of past research on model selection techniques in other statistical settings, algorithms for selecting the number of hidden layer nodes in a three layer, feedforward neural network are presented. The selection criterion attempts to "grow" the network beginning with a small initial number of hidden layer nodes (as opposed to pruning a relatively large network). For the nonlinear regression problem, the method is based on cross-validation estimates of the prediction mean prediction error for the candidate networks. For the classification problem, the method is based on resubstitution estimates of the misclassification probability for the candidate networks. Also considered is the use of principal components analysis on the training set in order to reduce the dimensionality of the input vector prior to "growing" the parsimonious network. Test cases and applications of the methods described herein will be included in a sequel (Part II), to be published separately, to illustrate the effectiveness of the methods.

## 1. INTRODUCTION

An artificial neural network (ANN) can be viewed as an analog computational device that implements a potentially highly nonlinear function. That is, an ANN simply computes the transfer function g in the relationship y=g(x), where g is a suitably well behaved (e.g. measurable, continuous, differentiable, etc.) mapping from the n-dimensional hypercube $[0,1]^n$ to the m-dimensional real numbers, $R^m$. A typical ANN, with five input nodes (neurons), three middle layer nodes, and two output layer nodes, is depicted in figure 1. A simplified explanation of the operation of this type of ANN, known as a single hidden layer feedforward ANN is described as follows.

The values of the components of an input vector x are "presented" to the "input layer nodes" of the network. Linear combinations of these values are formed using the "interconnection weights" $w_{ij}^{MI}$, and "fed forward" to the "middle layer nodes," each computing a function $F_m$, typically defined by the logistic "sigmoid" function $F(v) = \exp(v)(1+\exp(v))^{-1}$, on

its input. Linear combinations of these middle layer outputs are then formed using the interconnection weights $w_{ij}^{OM}$ and fed forward to the "output layer nodes" where a (typically, the same) function $F_o$ is applied to produce the final output y. Mathematically, the transfer function g is given by

$$y_i = F_o\left(\sum_j w_{ij}^{OM} F_m\left(\sum_k w_{jk}^{MI} x_k\right)\right) \tag{1}$$

where often, but not necessarily, $F_o=F_m=F$, $y_i$ is the ith component of y, and $x_k$ is the kth component of x. A powerful feature of such an ANN is its ability to approximate a wide variety of transfer functions g by varying the number of input, middle, and output layer nodes and the corresponding interconnection weights. In fact, it was proved (Kolmogorov, (1957)) that if g is continuous, then g has an exact representation of the type that could be implemented by a neural network of the type in figure 1, providing that the individual neurons be allowed to compute possibly different (not necessarily sigmoidal) transfer functions. In fact, this theory even specifies the number of middle layer nodes, (2n+1), if n is the number of input layer nodes (i.e. the dimension of the input vector x). Using the mathematical tools of functional analysis, it has been proven that, loosely speaking, fairly general functions g can be approximated to any desired degree of accuracy using the sigmoidal logistic function F for $F_m$ and $F_o$ in (1), and by increasing the number of middle layer nodes (see Cybenko (1989), for example).

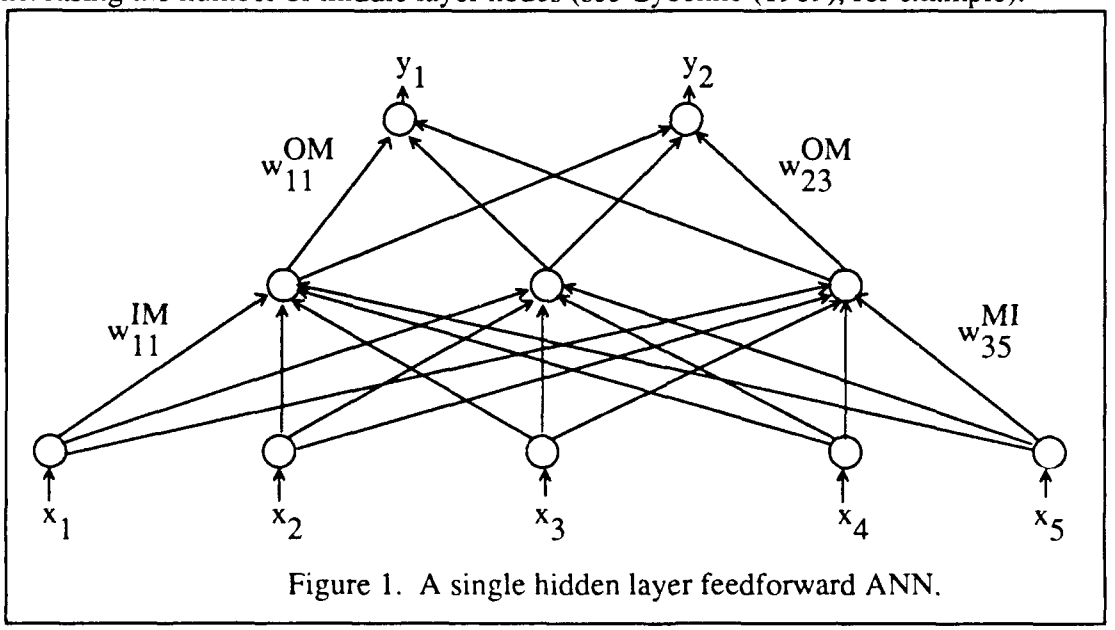



Figure 1. A single hidden layer feedforward ANN.

Because of this ability to approximate a wide range of multivariate functions, ANNs have been used as nonlinear regression functions for the purpose of developing predictive relationships. A formal mathematical model for nonlinear regression is

$$y = g(x; \theta) + \varepsilon \tag{2}$$

where y, x, and $\varepsilon$ are jointly distributed random m-, n-, and m-vectors, respectively, $\varepsilon$ is independent of y and x, $E(\varepsilon) = 0$, $Cov(\varepsilon) = E(\varepsilon\varepsilon^t) = \Sigma$, and $\theta$ is a vector of unknown parameters (the interconnection weights, in the case of an ANN). See, for example, Seber et.al. (1989) and Gallant (1987) for extensive treatments of models such as (2). One objective in using the model (2) is to first determine an estimate of the unknown parameter $\theta$ based on a random sample from the joint distribution of (x, y), and then use this estimate in the model to predict responses y at new inputs x. For ANNs, this procedure is carried out in principle by "training" the network on a sample of exemplars $(X_1, Y_1), ..., (X_N, Y_N)$, where $X_i$ and $Y_i$ are n- and m-dimensional vectors, respectively. An approach to this training is to determine the interconnection weights that minimize the mean squared error

$$Q = (1/N)\sum_{i=1}^{N}(Y_i-O_i)^t(Y_i-O_i) \tag{3}$$

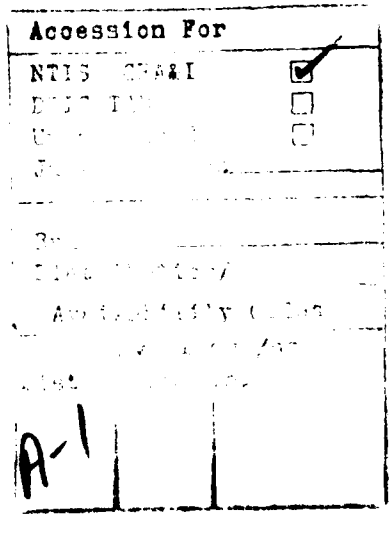
where $O_i$ is the actual output vector displayed on the output layer of the ANN when input vector $X_i$ is presented to the input layer, and "t" signifies matrix transpose, interpreting the vectors involved in (3) to be column vectors. Note that $O_i$ will generally differ from $Y_i$ due to random error (the "$\varepsilon$" in equation 2) and the approximation error (since the g in equation 2 may not be exactly of the parametric form implementable by an ANN; i.e. of the form of equation 1). The procedure (with variations) that is commonly used in determining the connection weights by minimizing (3) is the back-propagation algorithm (Soulie et.al. (1987), Hecht-Nielsen (1991), Hertz, et.al. (1991), and Rummelhart et.al. (1986)). A significant advantage of the back-propagation algorithm is that the network itself can carry out the minimization and estimation procedure, without external software, making it possible to implement a neural network completely in hardware or firmware. Having trained the network on the set of exemplars, it is then used to predict new responses based on new inputs. That is, the network is used to generalize.

Great success has been achieved in using neural networks in this fashion in many engineering, economic / financial, and biomedical applications. Recognition that the use of ANNs in this context fits within the framework of nonlinear regression analysis appeared in the literature fairly recently, although this was apparently understood by ANN researchers much earlier. Angus (1989) gives an exposition of this connection along with an interpretation of the back-propagation algorithm as a version of stochastic gradient descent. White's (1989) landmark paper was the first to show that in the context of statistical estimation theory, the back propagation estimator of the interconnection weights are relatively inefficient (i.e. have larger variance) compared to standard nonlinear least squares estimators, and presents a method (which amounts to taking one Newton-Raphson iteration step from the back-propagation estimators) for reducing the asymptotic variances of the back propagation estimators down to those of the nonlinear least squares estimators. Further along the lines of improving the statistical efficiency of back propagation estimators, Angus (1991) gives a Monte Carlo data generation method that reduces the mean squared error of the back propagation estimators when sufficient statistics are available.

Another use of ANNs that has shown promise, especially in the area of medical diagnosis, is in classification. Here, (X, Y) is assumed to be a random sample of size 1 from a probability distribution p(A, j), where A is a Borel subset of $R^n$, and $j \in \{1, ..., K\}$. Here, j is assumed to signify one of K distinct populations ("classifications", or diagnoses), and, conditional on Y=j, X has a distinct probability distribution on $R^n$. That is, $P\{X \in A, Y=j\} = p(A,j)$, and $P\{X \in A | Y=j\} = \int_A p(dx|j)$, where $p(A|j) = p(A,J)/\pi(j)$ and $\pi(j)$ is the marginal probability that Y=j (i.e. that "X comes from population j"). In the classification problem, X is observed (without its corresponding Y) and one must decide the population from which X came. That is, one must predict Y based on observing X. If $p(\bullet|j)$ has a density $f_j$ with respect to Lebesgue measure on $R^n$, and X=x is observed, then the optimum decision is to classify X into the population j for which $\pi(j|x)=f_j(x)\pi(j)\left[\sum_{j=1}^{K} f_j(x)\pi(j)\right]^{-1}$ is maximal. This latter quantity is $P\{Y=j|X=x\}$, and this decision rule is called the Bayes rule. In practice, neither the $\pi(j)$s nor the $f_j$s are known, but a training sample $(X_1, Y_1), ..., (X_N, Y_N)$ is available, and an approximation to the function $\pi(j|x)$ is "learned" by an appropriate ANN using, for example, back propagation

to minimize (3). Such an ANN would employ sigmoidal transfer functions at the output nodes that restrict the outputs to lie in [0,1] (such as the logistic sigmoid function), and the output layer would have m=K output nodes. The output Y=j would be designated by fixing the jth output node at 1, and the other nodes at 0. The network thus trained would then be used to classify new X values into one of the K populations. Other approaches to this classification problem include discriminant analysis, in which the Xs are assumed to come from one of K different multivariate normal populations, kernel density estimation, kth nearest neighbor rules, and classification / regression trees (CART). Breiman et.al. (1984) is the definitive reference for CART methods, and also gives a brief description of the first three methods. An ANN that directly attacks the kernel density estimation problem is studied by Marchette and Priebe (1989).

Despite its success in many sophisticated applications, this "training-generalization" application of ANNs has a serious drawback analogous to the misspecification problem (underfitting or overfitting) in classical statistical models, as exemplified in figure 2 for the nonlinear regression application. If the ANN architecture has enough middle layer nodes, then by forcing the mean squared error (3) to be small enough, the network can be made to collocate the exemplars exactly. Since the responses $Y_i$ in the exemplar set typically have measurement error according to the model (2), this means that the ANN can be made to force the approximating surface to pass through the points $(X_i, g(X_i;\theta)+\varepsilon_i)$ i=1, ..., N. This is what is meant by "fitting the noise," and it leads to a regression surface that is irregular (i.e. "bumpy") and hence poor at generalization. Similarly, if there are too few nodes in the middle layer, then the ANN will be a poor approximation to the true regression surface, being able to achieve only limited generalization capability. White (1981) discusses the misspecification problem for general models of the form (2).



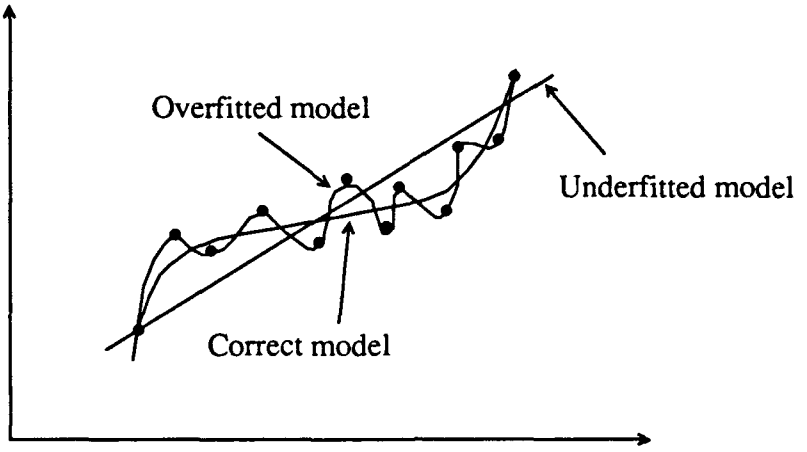



Figure 2. The effects of overfitting and underfitting.

Selection of the proper size of the network is thus of vital concern in applications, and is the major thrust of this paper. Following is a discussion of techniques that are used in other modeling contexts to select the proper parametric model, and general principles that aid in attacking the problem for ANNs.

## 2. APPROACHES AND PRINCIPLES IN MODEL SELECTION

Model selection has been studied extensively by many researchers for a variety of statistical models related to (2). For intrinsically linear models, selection is tantamount to selecting the proper regressor variables (e.g. linear, quadratic, cross product, etc.) along with the dimension of the unknown parameter space. For example, Mallows (1964, 1973) considers the general linear model, and the Mallows measure $C_p$ is extensively used in computerized stepwise linear regression packages. See also Myers (1990) for specific implementations and derivations of $C_p$. Model selection is also of extreme importance in time series models where prediction and/or interpolation are of interest. See, for example, Shibata (1976), Bhansali (1978), Hannan et.al. (1979), Wei (1987), and Hemerly et.al. (1989) for model selection criteria analysis for autoregressive and stochastic regression models under minimal distributional assumptions. Hemerly et.al. (1991) have also studied the problem of determining the order of an

autoregressive model when there is no a priori upper bound on this order.

When more specific distributional and parameter information are available (i.e. the distribution of the error vector $\varepsilon$ in (2), and prior information concerning $\theta$) several authors have found optimal model selection criteria in a Bayesian context. Among these, see Atkinson (1978) and Smith et.al (1980). Similarly, Schwarz (1978) has found a fairly simple criteria to apply in selecting a linear model from a subset of linear models (with bounded dimensions) that is asymptotically optimal when the observables follow a regular exponential family of distributions. Schwarz's criteria has been named the Schwarz Information Criteria (SIC), and will be discussed further later on. Other authors have studied the general problem of model selection based on various notions of information. See, for example, Akaike (1969, 1973, and 1974) (the "Akaike Information Criterion, or AIC), Stone (1977b, 1978), and Rissanen (1976, 1986).

A controlling theme in these investigations in model selection is the determination of a measure of prediction accuracy, or model fit, that takes into account both prediction variance and prediction bias, the former tending to increase with model complexity (e.g. the number of terms in the linear regression function), and the latter tending to decrease with model complexity. Based on a training sample, the "best" model is then selected to achieve some optimum balance between these competing quantities as estimated in some fashion from the sample. Naturally, the more information available concerning the distributional structure of the error term in (2), the more efficient is the model selection criterion in terms of accuracy and sample size necessary to achieve a decision. For example both the AIC and SIC (which is asymptotically optimal), require that the joint likelihood function of the observables be available and tractable in order to be used. In contrast, the PLS (Predictive Least Squares) criterion discussed in Hemerly et.al. (1989) for determining the order of an autoregressive model, and Mallows $C_p$ measure for linear models, are computable from relatively simple sample characteristics. By nature, however, all the above methods are computationally intensive, as they require fitting many candidate models to the training sample in order to make the final selection.

In the context of classification models, Breiman et.al. (1984) give extensive discussion and methods for growing and pruning classification and regression trees (CARTs) based on

cross-validation estimates of misclassification probabilities. There, the estimated misclassification probability (a measure of accuracy) is traded off against the number of terminal nodes of the tree (a measure of complexity).

Some specific work for ANNs has been done in the area of network size selection. These approaches generally fall into two categories: those that attempt to "prune" a large network of connections and/or nodes, and those that attempt to "grow" a larger network starting with a small network. The work in the latter category is mostly related to ANNs with nodes that take values in a discrete set (e.g. {0,1}, or {-1,1}). See, for example, Marchand et.al. (1990), Frean (1990), Mezard et.al. (1989), Sirat et.al. (1990), Fahlman et.al. (1990), and Gallant (1986). In the first category, Sietsma et.al. (1988), Hinton (1986), Scalettar et.al. (1988), Kramer et.al. (1989), Hanson et.al. (1989), and Chauvin (1989), have made contributions, attacking the problem by modifying the training rule (i.e. back propagation algorithm) to consider a penalty term in (3) to discourage complexity (complexity increases as number of interconnections and / or nodes increases). Other novel approaches to ANN selection and design include the use of genetic algorithms, which draw analogies with genetic natural selection for evolution and survival in biological populations (Miller et.al. (1989), and Harp et.al. (1990)).

These aforementioned approaches are either not suitable for continuum-valued input / output neurons, or they require extensive modification of the back propagation algorithm, a luxury that may not be feasible nor desirable (for example, if the ANN is implemented in a "canned" computer routine, or in hardware/firmware). In addition, they do not attempt to optimize the network with respect to some statistical measure of prediction error, and do not "preprocess" the input vector before presenting it to the ANN. The purpose of preprocessing would be to eliminate redundant information in the input vector, attempting to encompass most of the variability in the input sample space with far fewer dimensions. This concept is used successfully in human learning, the preprocessing initially accomplished via teachers and mentors, until the same level of input discrimination is learned by the pupil.

It is therefore proposed in this paper that the selection of an ANN for a given application and training sample be accomplished in two stages. The first stage involves preprocessing of the input data to achieve reduction in dimensionality if possible. The second stage is to grow an

appropriate ANN, without modifying the back propagation algorithm, that is of the "best" size in terms of balancing prediction variance with prediction bias (when regression structure is present as in (2)) or balancing misclassification probability with a measure of network complexity (when classification is the goal). More will be said concerning these tradeoffs later on. It is instructive at this point to review the principles behind some of these past techniques (e.g. of Schwarz (1978) and Mallows (1973)) for their pedagogic value and in order to motivate the algorithms that will be proposed later in this paper for the second stage of selection of an ANN. The first stage, the preprocessing of input data, will be accomplished via principal components analysis (see Rao (1973), for example). A relatively new technique, SIR (sliced inverse regression, Li (1991)), is also a feasible tool in achieving the dimensionality reduction of inputs in the case where the ANN yields 1-dimensional output. Further development of this concept will be a topic for future research.

Mallows' $C_p$ measure will now be reviewed briefly. For the moment, consider the full-rank linear model $Y=X\beta+\varepsilon$ where Y is an Nx1 observation vector, X is an Nxm matrix of "independent" variables, $\beta$ is a mx1 vector of unknown parameters, and $\varepsilon$ is an Nx1 error term that satisfies $E(\varepsilon)=0$, $Cov(\varepsilon)=\sigma^2 I$. It is well known that the least squares estimator of $\beta$ is $\hat{\beta} = (X^tX)^{-1}X^tY$, and this estimator is the best linear unbiased estimator of $\beta$. Suppose $0<p<m$, and that X and $\beta$ are partitioned as $X=(X_1|X_2)$, $\beta^t=(\beta_1^t|\beta_2^t)$, so that $Y=X_1\beta_1+X_2\beta_2+\varepsilon$ where $X_1$ is Nxp, $X_2$ is Nx(m-p), $\beta_1$ is px1, and $\beta_2$ is (m-p)x1, and that we fit an underspecified model by assuming that $\beta_2=0$, estimating $\hat{\beta}_1 = (X_1^tX_1)^{-1}X_1^tY$. Denote the rows of $X_1$ by $x_1^t, \ldots, x_N^t$. Then the predicted value of $y_i$, the $i^{th}$ component of Y, based on this fitted model, is $\hat{y}_i = x_i^t\hat{\beta}_1$. The total expected prediction error, summed over the data points and normalized by $\sigma^2$, is

$$E\left(\sum_{i=1}^N (\hat{y}_i-Ey_i)^2/\sigma^2\right) = \sum_{i=1}^N MSE(\hat{y}_i)/\sigma^2 = \sum_{i=1}^N \frac{var(\hat{y}_i) + bias^2(\hat{y}_i)}{\sigma^2}. \qquad (4)$$

Notice that $var(\hat{y}_i) = x_i^tCov(\hat{\beta}_1)x_i = \sigma^2 x_i^t(X_1^tX_1)^{-1}x_i$ and, since $\sum_{i=1}^N x_ix_i^t = X_1^tX_1$,

$$\sum_{i=1}^N x_i^t(X_1^tX_1)^{-1}x_i = \sum_{i=1}^N tr\left[(X_1^tX_1)^{-1}x_ix_i^t\right] = tr\left[(X_1^tX_1)^{-1}(X_1^tX_1)\right] = p,$$

where "tr" denotes matrix trace, so that (4) becomes

$$p + \sum_{i=1}^{N} \text{bias}^2(\hat{y}_i)/\sigma^2. \qquad (5)$$

Writing the vector of biases as $\hat{Y}-X\beta = X_1\hat{\beta}_1-X_1\beta_1-X_2\beta_2$, the sum in (5) can be written as

$$\sum_{i=1}^{N} \text{bias}^2(\hat{y}_i) = \left(E(X_1\hat{\beta}_1)-X_1\beta_1-X_2\beta_2\right)^t\left(E(X_1\hat{\beta}_1)-X_1\beta_1-X_2\beta_2\right)$$

$$=(X_2\beta_2)^t(I-X_1(X_1^tX_1)^{-1}X_1^t)X_2\beta_2, \qquad (6)$$

with (6) following since the matrix in the quadratic form is idempotent. Notice now that the expected value of the error sum of squares in fitting the underspecified model is given by

$$E(SSE) = E\left((Y-X_1\hat{\beta}_1)^t(Y-X_1\hat{\beta}_1)\right) = tr\left[E\left((I-X_1(X_1^tX_1)^{-1}X_1^t)YY^t\right)\right]$$

$$= tr\left[(I-X_1(X_1^tX_1)^{-1}X_1^t)(\sigma^2I+X\beta\beta^tX^t)\right] = \sigma^2(N-p) +(X_2\beta_2)^t(I-X_1(X_1^tX_1)^{-1}X_1^t)X_2\beta_2$$

$$= \sigma^2(N-p) + \sum_{i=1}^{N} \text{bias}^2(\hat{y}_i),$$

so that, letting $s^2$= SSE/(N-p), (4) becomes

$$p + (N-p)(E(s^2)-\sigma^2)/\sigma^2. \qquad (7)$$

If an independent estimate of $\sigma^2$ is available, call it $\hat{\sigma}^2$, then (7) can be estimated by the $C_p$ statistic (Mallows (1973)) given by

$$C_p = p +(N-p)(s^2-\hat{\sigma}^2)/\hat{\sigma}^2. \qquad (8)$$

The importance of the statistic (8) is that it expresses the tradeoff between the number of terms in the regression model (p) and the prediction bias, and it is used in selecting the best regression model by selecting the model with the lowest $C_p$ value among candidate models. This selection is usually carried out graphically by fitting the candidate models and plotting $C_p$ versus p, and selecting the model whose p is closest to the line $C_p$=p.

The PRESS (prediction sum of squares) method (see Myers (1991), ch. 4, for example) is based on the principle of considering the prediction error of a fitted model (i.e. fitted to a training sample) when used to predict the value of a new, independent exemplar. This principle also underlies the derivation of Akaike's (1974) AIC and (1969) FPE (final prediction error) criterion, as well as the PLS criterion studied by Hemerly et.al. (1989, 1991). The basic idea is to

define a measure of the prediction error for the fitted model with respect to a new, independent exemplar, and then develop a cross-validation estimator of the measure by splitting the training sample into a fitting sample and a validation sample. The PRESS method is described as follows.

Assume the same linear model setup as in the discussion of Mallows' $C_p$ statistic. For each i, i=1, ..., N, remove the pair $(x_i, y_i)$ from the training sample and fit the least squares estimator of $\beta$ based on the N-1 remaining points. Call the resulting estimator $\hat{\beta}_{-i}$, to designate that the ith data point has been removed. Form the prediction $\hat{y}_{i,-i} = x_i^t\hat{\beta}_{-i}$ of $y_i$, i=1, ..., N. Note that neither $x_i$ nor $y_i$ have been used in determining $\hat{\beta}_{-i}$. The PRESS residuals are defined by $e_{i,-i}= y_i-\hat{y}_{i,-i}$, i=1, ..., N, and the PRESS is defined to be

$$\text{PRESS}=\sum_{i=1}^{N} e_{i,-i}^2 .$$

PRESS contains both components of prediction variance and prediction bias, as does $C_p$. An advantage of the PRESS residuals is that they are particularly sensitive to points where prediction is poor, while ordinary residuals (which measure empirical fit) are not. In fact, it can be shown that the PRESS residuals are related to the ordinary residuals $e_i = y_i-x_i^t\hat{\beta}$ by the formula $e_{i,-i}=e_i/(1-x_i^t(X^tX)^{-1}x_i)$ as follows. Let $h_{ii}=x_i^t(X^tX)^{-1}x_i$. Then, by definition, the ith PRESS residual is

$$e_{i,-i} = y_i-x_i^t\left[X^tX-x_ix_i^t\right]^{-1}(X^tY-x_iy_i).$$

By the Sherman-Morrison-Woodbury Theorem (Rao (1973)),

$$\left[X^tX-x_ix_i^t\right]^{-1} = (X^tX)^{-1} + \frac{(X^tX)^{-1}x_ix_i^t(X^tX)^{-1}}{1-h_{ii}}$$

so that

$$e_{i,-i}= y_i-x_i^t(X^tX)^{-1}X^tY+x_i^t(X^tX)^{-1}x_iy_i-\frac{x_i^t(X^tX)^{-1}x_ix_i^t(X^tX)^{-1}X^tY}{1-h_{ii}}+\frac{x_i^t(X^tX)^{-1}x_ix_i^t(X^tX)^{-1}x_iy_i}{1-h_{ii}}$$

$$= \frac{y_i(1-h_{ii})-\hat{y}_i(1-h_{ii})+h_{ii}(1-h_{ii})y_i-h_{ii}\hat{y}_i+h_{ii}^2 y_i}{1-h_{ii}} = \frac{y_i-\hat{y}_i}{1-h_{ii}} = \frac{e_i}{1-h_{ii}}.$$

Thus, the PRESS can be computed without fitting the N-1 "leave-one-out" regressions by the formula

$$\text{PRESS} = \sum_{i=1}^{N}\left(\frac{e_i}{1-h_{ii}}\right)^2,$$

and the PRESS residuals are seen to be ordinary residuals, inflated by $(1-h_{ii})^{-1}$, where $h_{ii}=x_i^t(X^tX)^{-1}x_i$ is, apart from a missing factor of $\sigma^2$, the ordinary prediction variance. Hence, where prediction is poor (i.e. $h_{ii}$ close to 1), the PRESS residual is greatly inflated. When several candidate models are under consideration, the one with the smallest PRESS is the model of choice under this criterion.

Mallows' $C_p$ and the PRESS residual methods were derived under minimal distributional assumptions on the error vector $\varepsilon$, but with the fairly strong assumption that the regression was intrinsically linear. Essential use was made of this fact in computing $C_p$, and for deriving a simple computational formula for the PRESS statistic that avoids fitting all the linear regressions. When more specific information is available, the optimality of these selection procedures can be improved, and an (asymptotically) optir p.ocedure can be derived. This was accomplished by Schwarz (1978) as follows.

Suppose that $X_1, \ldots, X_N$ are a random sample from a regular exponential family of distributions with probability density, with respect to a $\sigma$-finite measure $\mu$ on the sample space, given by $f(x;\theta) = \exp(\theta^t y(x)-\eta(\theta))$, where $\theta$ ranges over the natural parameter space $\Theta$, a convex subset of the d-dimensional Euclidean space, and y is the sufficient d-dimensional statistic. The competing models are assumed to be defined by restricting the parameter space to subsets of the original $\Theta$ of the form $L_j\cap\Theta$, where each $L_j$ is a $d_j$-dimensional linear submanifold of d-dimensional Euclidean space, $0<d_j\leq d$, $j\in J$, where J is a finite index set. Suppose that $\theta$ has an a priori probability measure of the form $\tau\{d\theta\} = \sum_{j\in J}\alpha_j\tau_j\{d\theta\}$, where $\alpha_j=P\{\text{model j is correct}\}$,

and $\tau_j\{\bullet\}$, the conditional a priori distribution of $\theta$ given that model i is correct, has a nonsingular $d_j$-dimensional density on $L_j \cap \Theta$ with respect to $d_j$-dimensional Lebesgue measure that is bounded and locally bounded away from zero throughout $L_j \cap \Theta$. Notice that it is being assumed that the total number of possible models is finite, and that there is an upper bound, namely d, on the dimensionality of the model. In this Bayesian context, the optimum decision is to choose the model with the highest posterior probability. By Bayes' formula, the posterior probability that model j is correct, given $X_1, ..., X_N$, is

$$P\{\text{model j is correct} \mid X_1, ..., X_N\}$$

$$= \alpha_j \int_{L_j \cap \Theta} \exp(N(\theta^t Y - \eta(\theta))\tau_j\{d\theta\} \left[ \sum_{j \in J} \alpha_j \int_{L_j \cap \Theta} \exp(N(\theta^t Y - \eta(\theta))\tau_j\{d\theta\} \right]^{-1} \qquad (9)$$

where $Y=(1/N)\sum_{i=1}^{N} y(X_i)$. Since the normalizing constant is the same for each j, and since $x \mapsto \ln(x)$ is a monotone increasing mapping for $x>0$, the optimum decision is to choose the model corresponding to j having the largest value of

$$S(Y, N, j) = \ln(\alpha_j) + \ln\left( \int_{L_j \cap \Theta} \exp(N(\theta^t Y - \eta(\theta))\tau_j\{d\theta\} \right). \qquad (10)$$

The asymptotic behavior of (10) is of interest. An asymptotic expansion of (10) is easy to derive arguing heuristically. A rigorous derivation is given in Schwarz (1978).

From the theory of asymptotic expansions of integrals of the form in (10), it follows that the asymptotic behavior of the integral is determined by that of the integral over a small neighborhood about the value of $\theta$ at which $\theta^t Y - \eta(\theta)$ takes on its maximum in $L_j \cap \Theta$. Call this value $\theta_0$. Expanding $\theta^t Y - \eta(\theta)$ in a Taylor series about $\theta_0$, recognizing that the linear term in the expansion vanishes since $\theta_0$ yields a maximum, it follows that for $\theta$ near $\theta_0$, $\theta, \theta_0 \in L_j \cap \Theta$,

$$\theta^t Y - \eta(\theta) \approx \theta_0{}^t Y - \eta(\theta_0) - (1/2)(\theta - \theta_0)^t \Sigma_0^{-1} (\theta - \theta_0),$$

where $\Sigma_0^{-1} = \left( \frac{\partial^2 \eta(\theta_0)}{\partial \theta_0 \partial \theta_0^t} \right)$ is nonnegative definite, since it is the covariance matrix of $y(X_1)$ when

$\theta=\theta_0$. Using this in (10) gives

$$S(Y, N, j) \sim N \sup_{\theta \in L_j \cap \Theta}(\theta^t Y - \eta(\theta)) - \ln\left( \int_{L_j \cap \Theta} e^{-N(1/2)(\theta-\theta_0)^t \Sigma_0^{-1}(\theta-\theta_0)} \tau_j\{d\theta\} \right)$$

Let $f_j$ be the density of $\tau_j$. Because of the assumptions on $f_j$, and by making a linear transformation of the integration variables, the last written integral is asymptotic to

$$f_j(\theta_0) \int_{R^{d_j}} e^{-N(1/2)\lambda \|\varphi\|^2} d\varphi = f_j(\theta_0)\,(2\pi)^{d_j/2} \left(\frac{1}{N\lambda}\right)^{d_j/2},$$

where $\lambda$ is a positive constant. Using this in (10) yields

$$S(Y, N, j) \sim N \sup_{\theta \in L_j \cap \Theta}(\theta^t Y - \eta(\theta)) - (d_j/2)\ln(N), \text{ as } N \to \infty. \tag{11}$$

Notice that the supremum in (11) is just the maximum of the log-likelihood, the maximum being taken over $L_j \cap \Theta$. Denoting the maximum over $\theta \in L_j \cap \Theta$ of the likelihood function by $M_j(X_1,...,X_N)$, the Schwarz criterion is, from (11), to choose the model j for which

$$\ln(M_j(X_1, ..., X_N)) - (d_j/2)\ln(N) \tag{12}$$

is maximized, $j \in J$.

Akaike (1974) defines a similar criterion to (12), called the AIC, which amounts to choosing the model j having the maximum value of

$$\ln(M_j(X_1, ..., X_N)) - d_j. \tag{13}$$

Of course, the work of Schwarz (1978) and the preceding discussion shows that (13) is not asymptotically optimum in the aforementioned setting. Because of the ln(N) multiplier in (12), the SIC tends to favor lower dimensional models than the AIC, and in fact, several authors have observed that the AIC tends to overestimate the dimension of the model (Shibata (1976), Jones (1975)).

Despite the pedagogic value of these considerations, it is clear that neither Mallows' $C_p$ nor the SIC are directly applicable to the determination of the best size for a neural network based on a given set of training data. Conceptually, however, they suggest that a good procedure

would be based on selecting a network that achieves a balance between the closeness with which the model fits the training data, and the "dimensionality" of the approximating regression surface.

## 3. ANN SELECTION FOR THE NONLINEAR REGRESSION PROBLEM

In this section an algorithm is described and proposed for selecting the size of a single hidden layer feedforward neural network as described in section 1 and figure 1. The "size" of the network will be defined to be the number of hidden layer nodes, the input layer and output layer sizes being dictated by the dimensions of the $X_i$ (input) and $Y_i$ (output) vectors. No attempt is being make here to limit the number of interconnections for a given network size. This section treats the case in which regression structure is present.

Suppose, as in section 1, that a random training sample $(X_1, Y_1), \ldots, (X_N, Y_N)$ is available. Here, it is assumed that $X_i$ and $Y_i$ are jointly distributed *n*- and *m*- dimensional random vectors, respectively. Suppose that $Y_i = g_p(X_i;\theta) + \varepsilon_i$, where $E(\varepsilon_i)=0$, $cov(\varepsilon_i)=\Sigma$, i=1, ..., N, and that the $\varepsilon_i$s are independent and identically distributed. The function $g_p$ is assumed to belong to the class of functions that are represented by a single hidden layer feedforward neural network with p hidden nodes and given node transfer functions (e.g. logistic sigmoid functions). Thus, $g_p$ depends also on the (m+n)p interconnection weights of the network, represented by the vector $\theta$. To be definite, assume that the back propagation algorithm is used in fitting the interconnection weights for a given value of p. This assumption is not essential, as any suitable numerical method for finding the weights based on a given training sample will suffice.

For an ANN of size p, assume that the weights have been fit based on the training sample, and that a new exemplar (X, Y), independent of the training set $T=((X_1,Y_1), \ldots, (X_N,Y_N))$, is available. As usual, it is assumed that T constitutes a random sample from the joint distribution of (X,Y). Define the prediction mean squared error by

$$PMSE = E\left(\| g_p(X;\hat{\theta}) - g_{p_0}(X;\theta) \|^2\right). \qquad (14)$$

where $p_0$ is the correct size for the network, $\theta$ is the true value of the weight vector, and $\|\bullet\|$ is the Euclidean norm on $R^m$. In (14), $g_p(X;\hat{\theta})$ is the predicted value of Y based on knowledge of X, using the ANN that implements the transfer function $g_p(\bullet;\hat{\theta})$. The hat "^" indicates that the interconnection weights in $g_p$ have been estimated from the training sample, so that $g_p$ contains the training sample information through its dependence on these estimated weights. The value $g_{p_0}(X;\theta)$ is the expected value of Y given X, since $g_{p_0}(\bullet;\theta)$ is the true regression function. The prediction mean squared error (14) contains both prediction variance and bias components analogous to (4) in the derivation of Mallows' $C_p$. In fact, ("tr" indicates matrix trace)

$$\text{PMSE} = E\,tr\Big[\text{Cov}\big(g_p(X;\hat{\theta})|X\big)\Big] + E\,tr\Big[\text{Bias}\big(g_p(X;\hat{\theta})|X\big)\text{Bias}^t\big(g_p(X;\hat{\theta})|X\big)\Big] \equiv \wp_N(p),$$

the latter notation used to indicate the dependence on both N and p, and if m=1 (i.e. the response variable Y is 1-dimensional), then

$$\text{PMSE} = E\Big(\text{Var}(g_p(X;\hat{\theta})|X)\Big) + E\Big(\text{Bias}^2(g_p(X;\hat{\theta})|X)\Big).$$

Unfortunately, $g_{p_0}$ is unknown, and there are no tractable analytical calculations, as in the case of $C_p$, that render (14) useable for estimating p. The technique of ordinary cross-validation (OCV) estimation can be used to estimate (14). This method, also known as the leave-one-out method, is implemented as follows.

For each $k \in \{1, ..., N\}$, remove the exemplar $(X_k, Y_k)$ from the training set, and train the network on the remaining N-1 exemplars. Let $\hat{g}_p^{(k)}$ be the resulting estimated transfer function. Note that $\hat{g}_p^{(k)}$ is statistically independent of $(X_k, Y_k)$. Conditional on $X_k$, $Y_k$ is an unbiased estimator of $g_{p_0}(X_k;\theta)$. Hence, the OCV estimator of (14) is

$$V(p) = (1/N)\sum_{k=1}^{N} \|\hat{g}_p^{(k)}(X_k) - Y_k\|^2. \qquad (15)$$

The OCV estimator V(p) is biased in estimating E(PMSE). In fact,

$$E(V(p)) = \wp_{N-1}(p) + tr(\Sigma)$$

but the term $tr(\Sigma)$ is a constant, so that V(p) is still a measure of prediction bias plus variance, the former tending to increase for p decreasing away from $p_0$, while the latter tending to increase for p increasing away from $p_0$. Hence, a reasonable estimate of $p_0$ based on the training sample is

$$\hat{p}_0 = \arg\min V(p), \tag{16}$$

that is, the value of p that minimizes V(p) with respect to p.

OCV estimators of various prediction figures of merit have been extensively and successfully used in many contexts. See, for example, Breiman et.al. (1984) and Breiman (1991) for its use in classification and regression trees and regression splines, Wahba (1990) for its use in spline models for observational data, and Myers (1990, ch. 4) for its use in selecting a regression function and a comparison with Mallows' $C_p$ criterion. Theoretical work concerning cross-validation estimation has been carried out by Stone (1974), (1977a) and (1977b).

Use of (16) in this fashion requires, of course, that potentially many values of V(p) be computed, for various values of p, in order to locate the minimum value and corresponding p. Convergence of the weight estimation algorithm (i.e. the back propagation algorithm in this case) in computing the predictors $\hat{g}_p^{(k)}$ for each fixed p and k varying from 1 to N should be fairly rapid, as the weights computed from the previous value of k can be used as starting values for the algorithm for the next k. Nevertheless, this procedure is computationally intensive, and it is imperative that an attempt be made to reduce the dimensionality of the input vector prior to attempting to determine the best ANN via (16). Approaches to this are discussed later on.

## 4. ANN SELECTION FOR THE CLASSIFICATION PROBLEM

In this section, ANN selection is considered in the context of the classification problem. Here, the training set of exemplars $T=((X_1,Y_1), ..., (X_N,Y_N))$ constitutes a random sample from the distribution $p(A,j) = P\{X \in A, Y=j\}$ where A is a Borel set in $R^n$, and $j \in \{1, ..., K\}$ represents K distinct populations. The interpretation of an examplar $(X_k,Y_k)$ is that the measurement variable $X_k$ was generated from a subject in the population $Y_k$. The ANN is thus being used to estimate the function $\pi(j|x) = P\{Y=j|X=x\}$.

As described in section 3, an ANN of size p will constitute a three layer, single hidden layer ANN with p nodes in the middle layer, n input nodes, and m=K output nodes. The output neurons will be assumed to implement a sigmoidal transfer function that guarantees that each output node outputs a value in [0,1]. The output Y=j will be presented to the network during training by fixing the jth output node at 1, and all other output nodes at 0. A nontraining output from the ANN will be judged to constitute "Y=j" if the jth output node has the largest output value. This convention is made in lieu of constraining the output node values to sum to 1.

As in section 3, it will be assumed that the ANNs are trained based on T using the back propagation algorithm, and that no attempt is made to limit the number of connections ("zero out" weights) within a network of a given size.

Suppose a network of size p has been trained on the set T, yielding the decision rule $\hat{d}_p$. That is, $\hat{d}_p$ is a mapping from $R^n$ to {1,...,K} with the interpretation that if X=x is observed, then classify X into population $\hat{d}_p(x)$. By the previous discussion, $\hat{d}_p(x) = j$ if, when presented with input x, the ANN output node j produces the largest value over all output nodes. Let a new sample become available, (X, Y). The misclassification probability $MP^*$ is defined by

$$MP^*(N,p) = P\{\hat{d}_p(X) \neq Y | T\}, \tag{17}$$

the conditional probability (conditional on the training sample T) that the rule $\hat{d}_p$ fails to correctly classify the new sample. Since the joint distribution of (X,Y) is unknown, (17) cannot be computed. The resubstitution estimate of (17) is

$$MP(N,p) = (1/N)\sum_{i=1}^{N} I\{\hat{d}_p(X_i) \neq Y_i\}. \tag{18}$$

In (18), I{S} =1 if S is true, and I{S}=0 otherwise. Notice that in (18), the rule $\hat{d}_p$ is determined with the same data used to estimate the probability of misclassification. Therefore, MP in (18) is likely to give overly optimistic estimates of (17), and would therefore not be appropriate by itself as a figure of merit in sizing the ANN.

Drawing analogy with the work of Breiman et.al. (1984) for CARTs, and borrowing their notation and terminology, define the cost-complexity measure

$$R_\alpha(N,p) = MP(N,p) + \alpha p, \tag{19}$$

where $\alpha \geq 0$ is called the complexity parameter. Because of overfitting to the training data, MP will tend to decrease as p, the size (complexity) of the ANN, increases. Conversely, MP will tend to increase as p decreases. Thus, assuming that the true conditional probability function $\pi(j|x)$ belongs to the parametric family of functions that are implemented by the ANN, the ANN true size $p_0$ can be estimated by the value that minimizes (19). That is, the selection rule would be to choose the network size $\hat{p}_0$ such that

$$\hat{p}_0 = \arg\min R_\alpha(N,p). \tag{20}$$

At present, there are no guidelines for choosing $\alpha$, the complexity cost per hidden layer node in (20). This will be a topic of study in the sequel (Part II) to this study.

## 5. PREPROCESSING OF THE TRAINING SAMPLE

Let $(X_1, Y_1), \ldots, (X_N, Y_N)$ be the training sample as defined in section 3. Typically, the dimension of the $X_i$s, namely n, is fairly large. Moreover, it is often the case that there is strong linear association between some of the components of the Xs (e.g. pulse rate and respiration rate) in which case, some of the information contained in the X vector may be redundant. In order to eliminate some of this redundancy and, in effect, reduce the dimensionality of the $X_i$s, the following method, called principal components analysis, can be used. This principal components technique is successfully used to eliminate multicolinearity in linear regression models (see Myers (1991), ch. 8, for example).

Let $\hat{\Sigma}$ be the sample covariance matrix of the $X_i$s, given by $\hat{\Sigma}=(1/N)\Sigma_{i=1}^{N}(X_i-\bar{X})(X_i-\bar{X})^t$, $\bar{X}=(1/N)\Sigma_{i=1}^{N}X_i$. Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \geq 0$ be the ordered eigenvalues of $\hat{\Sigma}$ and $v_1, \ldots, v_n$ the corresponding set of orthonormal eigenvectors. Fix a threshold $\gamma \in (0,1)$ (typically, $\gamma$=.9 or .95) and select $n_0<n$ to be the smallest integer such that $\Sigma_{i=1}^{n_0}\lambda_i \geq \gamma\Sigma_{i=1}^{n}\lambda_i$. Let V be the $n_0 x n$ matrix whose $i^{th}$ row is given by $v_i^t$. Let $Z_i=VX_i$, i=1, ..., N. Each $Z_i$ represents a "reduced" version of $X_i$, the extraneous $n-n_0$ dimensions being eliminated because they are associated with small

eigenvalues of $\hat{\Sigma}$, which are in turn associated with high degrees of multicolinearity (redundancy) in the components of the $X_i$s. The new training sample to be used in the ANN is now given by $(Z_1, Y_1), ..., (Z_N, Y_N)$.

## 6. SUMMARY AND CONCLUSIONS

Promising approaches have been presented to choosing the best size of a single hidden layer feedforward ANN in both the nonlinear regression context as well as the classification context. Although restriction to this type of ANN architecture has been assumed, it is not viewed as a limitation in applications since there is compelling theoretical evidence that such an architecture maintains sufficient potential for functional approximation. These approaches have been designed to be applicable to an ANN without modifying its training rule (e.g. back propagation) or basic architecture (i.e. single hidden layer feedforward type). Other approaches, whereby the ANN size selection is embedded into the learning algorithm itself, were not addressed in this investigation. The effectiveness of the methods proposed will be the topic of the sequel to this report, Part II, in which simulation examples will be used to determine if the methods select the correct (or nearly correct) size of ANNs in both the nonlinear and classification contexts.

In order to render the proposed size selection algorithms more computationally feasible, an approach has been proposed for reducing the dimensionality of the input data to an ANN. This approach, the principal components approach, has been successful in classical statistical models having similar structure, and in many types of applications it effectively eliminates strong multicolinearities in high dimensional input data vectors (see Myers (1990), ch. 8, section 4, and Press (1981), ch. 9, section 4, for example). This application of principal components to ANNs will often greatly reduce the potential size of the ANN, thereby reducing the computation time entailed in applying the size selection algorithms presented herein.

The size selection criteria presented herein have been chosen based on proven principles in other statistical modeling problems (e.g. CARTs, time series models, and regression models). There are many factors that will have an effect on their ultimate performance in applications, however. For example, it has been assumed that back propagation will be used to train the

ANNs in applications. When using this method, the error threshold in minimizing (3) is typically selected by the user, and will have an effect on the ANN weights. The sensitivity of the size selection criteria presented here to this error threshold will need to be investigated. Other factors that require investigation include selection of the initial size of network for a given problem, the selection of the complexity cost per hidden layer node parameter in the classification problem, and the statistical properties of the proposed size estimators, $\hat{p}_0$.

Finally, the criteria proposed here may perform well in simulation examples where the data are actually generated via a nonlinear function of the type implemented by an ANN (i.e. of the parametric form given by equation 1). In reality ANNs produce, at best, approximations to naturally occurring functions. These naturally occuring functions are generally not of the exact parametric form of those that can be implemented by an ANN (see (1)), and hence they do not have their own "true" $p_0$ values associated with them (recall that the definition of $p_0$ on page 16 tacitly assumes that the unknown function is actually of the parametric form of equation 1). Nevertheless, finding the best size $p_0$ for an approximating ANN is often still achievable. Indeed, this difficulty (i.e. that of the underlying natural model not being of the a priori assumed parametric form) pervades statistical modeling in general, and yet "idealized," parsimoniously determined parametric models continue to provide useful and important answers to quantitatively posed questions. Therefore, further investigations into the effectiveness (and other open questions) of the selection criteria for simulated and real data will be important and worthwhile undertakings.

## REFERENCES


Akaike, H. (1969). Fitting autoregression models for prediction. Ann. Inst. Statist. Math. 21, pp. 243-247.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. 2nd International Symposium on Information Theory, pp. 267-281. Budapest: Akademia Kiado.

Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control AC-19, pp. 716-723.

Angus, J.E. (1989). On the connection between neural network learning and multivariate nonlinear least squares estimation. The International Journal of Neural Networks, vol. 1, no. 1., pp. 42-47.

Angus, J.E. (1991). Computer-assisted improvement of the estimation mean squared error with application to back propagation neural networks. Unpublished, under editorial review.

Atkinson, A.C. (1978). Posterior probabilities for choosing a regression model. Biometrika 65, 39-48.

Bhansali, R.J. and Downham, D.Y. (1977). Some properties of the order of an autoregressive model selected by a generalization of Akaike's EPF criterion. Biometrika 64, pp. 547-551.

Breiman, L. (1991). The Π method for estimating multivariate functions from noisy data. Technometrics 33(2), pp. 125-143 (with discussion).

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C. J. (1984). Classification and Regression Trees. Monterey: Wadsworth and Brooks/Cole.

Chauvin, Y. (1989). A back-propagation algorithm with optimal use of hidden units. In Advances in Neural Information Processing Systems I (Denver 1988), ed. D.S. Touretzky, pp. 519-526. San Mateo: Morgan Kaufmann.

Cybenko, G. (1989). Approximation by superposition of a sigmoidal function. Mathematics of Control, Signals, and Systems 2, pp. 303-314.

Durrett, R. (1990). Probability: Theory and Applications. Belmont, CA: Brooks/cole.

Fahlman, S.E. and Lebiere, C. (1990). The cascade-correlation learning architecture. In Advances in Neural Information Processing Systems II (Denver 1989), ed. D.S. Touretzky, pp. 524-532. San Mateo: Morgan Kaufmann.

Frean, M. (1990). The upstart algorithm: a method for constructing and training feedforward neural networks. Neural Computation 2, 198-209.

Gallant, A.R. (1987). Nonlinear Statistical Models. New York: John Wiley and Sons.

Gallant, S.I. (1986). Optimal linear discriminants. In Eighth International Conference on Pattern Recognition (Paris 1986), pp. 849-852. New York: IEEE.

Hannan, E.J. and Quinn, B.G. (1979). The determination of the order of an autoregression. Journal of the Royal Statistical Society B 41(2), pp. 190-195.

Hanson, S.J. and Pratt, L. (1989). A comparison of different biases for minimal network construction with back-propagation. In Advances in Neural Information Processing Systems I (Denver 1988), ed. D.S. Touretzky, pp. 177-185. San Mateo: Morgan Kaufmann.

Harp, S.A., Samad, T., and Guha, A. (1990). Designing application-specific neural networks using genetic algorithm. In Advances in Neural Information Processing Systems II (Denver 1989), ed. D.S. Touretzky, pp. 447-454. San Mateo: Morgan Kaufmann.

Hecht-Nielsen, R. (1991). Neurocomputing. Reading, Mass.: Addison Wesley.

Hemerly, E.M. and Davis, M.H.A. (1989). Strong consistency of the PLS criterion for order determination of autoregressive processes. Annals of Statistics 17(2), pp. 941-946.

Hemerly, E.M. and Davis, M.H.A. (1991). Recursive order estimation of autoregressions without bounding the model set. Journal of the the Royal Statistical Society B 53(1), pp. 201-210.

Hertz, J., Krogh, A., and Palmer, R.G. (1991). Introduction to the Theory of Neural Computation. Reading, Mass.: Addison Wesley.

Hinton, G.E. (1986). Learning distributed representations of concepts. In Proceedings of the Eighth Annual Conference of the Cognitive Science Society (Amherst 1986), pp. 1-12. Hillsdale: Erlbaum.

Jones, R.H. (1975). Fitting autoregressions. Journal of the American Statistical Association 70, pp. 590-592.

Kolmogorov, A.N. (1957). On the representation of continuous functions of many variables by superposition of continuous functions of one variable and addition. [in Russian], Dokl Akad. Nauk USSR 114, pp. 953-956.

Kramer, A.H. and Sangiovanni-Vincentelli, A. (1989). Efficient parallel learning algorithms for neural networks. In Advances in Neural Information Processing Systems I (Denver 1988), ed. D.S. Touretzky, pp. 40-48. San Mateo: Morgan Kaufmann.

Li, K.C. (1991). Sliced inverse regression for dimension reduction. Journal of the American Statistical Association 86(414), pp. 316-327 (with discussion).

Mallows, C.L. (1964). Choosing variables in a linear regression: a graphical aid. Presented at the Central Regional Meeting of the Institute of Mathematical Statistics, Manhattan, Kansas.

Mallows, C.L. (1973). Some comments on $C_p$. Technometrics 15, pp. 661-675.

Marchette, D.J. and Priebe, C.E. (1989). The adaptive kernel neural network. Technical document 1676, Naval Ocean Systems Center, San Diego, CA 92152-5000.

Marchand, M., Golea, M., and Rujan, P. (1990). A convergence theorem for sequential learning in two layer perceptrons. Europhysics Letters 11, pp. 487-492.

Mezard, M. and Nadal, J.-P. (1989). Learning in feedforward layered networks: the tiling algorithm. Journal of Physics A 22, 2191-2204.

Miller, G.F., Todd, P.M., and Hegde, S.U. (1989). Designing neural networks using genetic algorithms. In Proceedings of the Third International Conference on Genetic Algorithms (Arlington 1989), ed. J.D. Schaffer, pp. 379-384. San Mateo: Morgan Kaufmann.

Myers, R.H. (1990). Classical and Modern Regression with Applications. Boston: PWS-Kent.

Press, S.J. (1981). Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference. Malabar, Florida: Robert E. Krieger.

Rao, C.R. (1973). Linear Statistical Inference and its Applications, 2nd ed. New York: John Wiley & Sons.

Rissanen, J. (1976). Modeling by shortest data description. Automatica 14, pp. 465-471.

Rissanen, J. (1986). Stochastic complexity and modeling. Annals of Statistics 14(3), pp. 1080-1100.

Rummelhart, D.E. and McClelland, J.L. and the PDP Research Group (1986). Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol. 1. Cambridge: MIT Press.

Scalettar, R. and Zee, A. (1988). Emergence of grandmother memory in feed forward networks: learning with noise and forgetfulness. In Connectionist Models and Their Implications: Readings from Cognitive Science, eds. D. Waltz and J.A. Feldman, pp. 309-332. Norwood: Ablex.

Schwarz, G. (1978). Estimating the dimension of a model. Annals of Statistics 6(2), pp. 462-464.

Seber, G.A.F. and Wild, C.J. (1989). Nonlinear Regression. New York: John Wiley and Sons.

Shibata, R. (1976). Selection of the order of an autoregressive model by Akaike's information criterion. Biometrika 63, pp. 117-126.

Sietsma, J. and Dow, R.J.F. (1988). Neural net pruning - why and how. In IEEE International Conference on Neural Networks (San Diego, 1988), vol. I, pp. 325-333. New York: IEEE.

Sirat, J.-A., and Nadal, J.-P. (1990). Neural trees: a new tool for classification. Preprint, Laboratoires d'Electronique Philips, Limeil-Brevannes, France.

Smith, A.F.M. and Spiegelhalter, D.J. (1980). Bayes factors and choice criteria for linear models. Journal of the Royal Statistical Society 42(2), pp. 213-220.

Soulie, F.F., Robert, Y. and Tchuente, M., eds. (1987). Automata Networks in Computer Science: Theory and Applications. Princeton: Princeton University Press.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society B 36, pp. 111-147.

Stone, M. (1977a). Asymptotics for and against cross-validation. Biometrika 64, pp. 29-35.

Stone, M. (1977b). An asymptotic equivalence of choice model by cross-validation and Akaike's criterion. Journal of the Royal Statistical Society B 39, pp. 44-47.

Stone, M. (1978). Comments on model selection criteria of Akaike and Schwarz. Journal of the Royal Statistical Society B 41, pp. 276-278.

Wahba, G. (1990). Spline Models for Observational Data. Philadelphia: SIAM.

Wei, C.Z. (1987). Adaptive prediction by least squares predictors in stochastic regression models with applications to time series. Annals of Statistics 15(4), pp. 1667-1682.

White, H. (1984). Asymptotic Theory for Econometricians. New York: Academic Press.

White, H. (1981). Consequences and detection of misspecified nonlinear regression models. Journal of the American Statistical Association 76(374), pp. 419-433.

White, H. (1989). Some asymptotic results for learning in single hidden-layer feedforward network models. Journal of the American Statistical Association, vol. 84, no. 408, pp. 1003-1013.

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE

# REPORT DOCUMENTATION PAGE

| | |
|---|---|
| 1a. REPORT SECURITY CLASSIFICATION<br>Unclassified | 1b. RESTRICTIVE MARKINGS<br>N/A |
| 2a. SECURITY CLASSIFICATION AUTHORITY<br>N/A | 3. DISTRIBUTION / AVAILABILITY OF REPORT<br>Approved for public release; distribution unlimited. |
| 2b. DECLASSIFICATION / DOWNGRADING SCHEDULE<br>N/A | |
| 4. PERFORMING ORGANIZATION REPORT NUMBER(S)<br>Report No. 91- 16 | 5. MONITORING ORGANIZATION REPORT NUMBER(S) |
| 6a. NAME OF PERFORMING ORGANIZATION<br>Naval Health Research Center<br>6b. OFFICE SYMBOL (*If applicable*)<br>Code 22 | 7a. NAME OF MONITORING ORGANIZATION<br>Chief, Bureau of Medicine and Surgery |
| 6c. ADDRESS (*City, State, and ZIP Code*)<br>P. O. Box 85122<br>San Diego, CA 92186-5122 | 7b. ADDRESS (*City, State, and ZIP Code*)<br>Navy Department<br>Washington, DC 20372-5120 |
| 8a. NAME OF FUNDING / SPONSORING ORGANIZATION Naval Medical Research & Development Command<br>8b. OFFICE SYMBOL (*If applicable*) | 9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER<br>American Socieity for Engineering Education (ASEE) Navy Summer Faculty Resarch Program |
| 8c. ADDRESS (*City, State, and ZIP Code*)<br>NNMC<br>Bethesda, MD 20889-5044 | 10. SOURCE OF FUNDING NUMBERS<br>PROGRAM ELEMENT NO. / PROJECT NO. / TASK NO / WORK UNIT ACCESSION NO. |

11. TITLE (*Include Security Classification*)
CRITERIA FOR CHOOSING THE BEST NEURAL NETWORK: PART I

12. PERSONAL AUTHOR(S) Angus, J.E., Ph.D.

| 13a. TYPE OF REPORT | 13b. TIME COVERED | 14. DATE OF REPORT (*Year, Month, Day*) | 15. PAGE COUNT |
|---|---|---|---|
| FINAL | FROM _____ TO _____ | 91 July 24 | 27 |

16. SUPPLEMENTARY NOTATION

| 17. COSATI CODES | | | 18. SUBJECT TERMS (*Continue on reverse if necessary and identify by block number*) |
|---|---|---|---|
| FIELD | GROUP | SUB-GROUP | Regression, Classification, Overfitting, Underfitting, Principal Components |
| | | | |
| | | | |

19. ABSTRACT (*Continue on reverse if necessary and identify by block number*)

**An investigation into the problem of determining a parsimonious neural network for use in prediction/generalization based on a given fixed learning sample was undertaken. Both the classification and nonlinear regression contexts were addressed. An exposition and survey of the problem and past research on model selection techniques in other statistical settings was compiled, and algorithms for selecting the number of hidden layer nodes in a three layer, feedforward neural network were developed. The selection criteria developed attempt to "grow" the networks beginning with a small initial number of hidden layer nodes (as opposed to pruning a relatively large network). For the nonlinear regression problem, the method is based on cross-validation estimates of the prediction mean squared error for the candidate networks. For the classification problem, the method is based on a cost complexity measure of the candidate networks based on resubstitution estimates of the probability of misclassification and a penalty function of the number of hidden layer nodes. Also considered was the use of principal**

| | |
|---|---|
| 20. DISTRIBUTION / AVAILABILITY OF ABSTRACT<br>☐ UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT ☐ DTIC USERS | 21. ABSTRACT SECURITY CLASSIFICATION<br>Unclassified |
| 22a. NAME OF RESPONSIBLE INDIVIDUAL<br>William Pugh | 22b. TELEPHONE (*Include Area Code*) 619-553-8403<br>22c. OFFICE SYMBOL<br>Code 22 |



★U.S. Government Printing Office: 1985—607-047